Emerging Trends in Drugs, Addictions, and Health 5 (2025) 100170

Contents lists available at ScienceDirect

Emerging Trends in Drugs, Addictions, and Health

journal homepage: www.elsevier.com/locate/etdah

Introducing CounseLLMe: A dataset of simulated mental health dialogues for comparing LLMs like Haiku, LLaMAntino and ChatGPT against humans

Edoardo Sebastiano De Duro, Riccardo Improta, Massimo Stella *

CogNosco Lab, Department of Psychology and Cognitive Science, Università degli Studi di Trento, Italy

ARTICLE INFO

Keywords:
Natural Language Processing
Mental Health
Large Language Models

ABSTRACT

We introduce CounseLLMe as a multilingual, multi-model dataset of 400 simulated mental health counselling dialogues between two state-of-the-art Large Language Models (LLMs). These conversations - of 20 quips each - were generated either in English (using OpenAI's GPT 3.5 and Claude-3's Haiku) or Italian (with Claude-3's Haiku and LLaMAntino) and with prompts tuned with the help of a professional in psychotherapy. We investigate the resulting conversations through comparison against human mental health conversations on the same topic of depression. To compare linguistic features, knowledge structure and emotional content between LLMs and humans, we employed textual forma mentis networks, i.e. cognitive networks where nodes represent concepts and links indicate syntactic or semantic relationships between concepts in the dialogues' quips. We find that the emotional structure of LLM-LLM English conversations matches the one of humans in terms of patient-therapist trust exchanges, i.e. 1 in 5 LLM-LLM quips contain trust along 10 conversational turns versus the 24% rate found in humans. ChatGPT and Haiku's simulated English patients can also reproduce human feelings of conflict and pessimism. However, human patients display non-negligible levels of anger/frustration that is missing in LLMs. Italian LLMs' conversations are worse in reproducing human patterns. All LLM-LLM conversations reproduced human syntactic patterns of increased absolutist pronoun usage in patients and second-person, trust-inducing, pronoun usage in therapists. Our results indicate that LLMs can realistically reproduce several aspects of human patient-therapist conversations and we thusly release CounseLLMe as a public dataset for novel data-informed opportunities in mental health and machine psychology.

1. Introduction

Large Language Models (LLMs) are Artificial Intelligences (AIs) trained to produce human-like texts. While LLMs were born as neural networks in the realm of computer science (Basile et al., 2023), they are quickly becoming more and more investigated across multidisciplinary frameworks outside of informatics (Stella et al., 2023). The ever-growing field of machine psychology (Abramski et al., 2023; Hagendorff, 2023) deals with exploring the cognitive abilities of LLMs in terms of associating ideas for representing reasoning, perceptual and emotional processes (Bertolazzi et al., 2023; Raz et al., 2024). LLMs can indeed perform complex tasks relative to acquiring and producing knowledge, with important applications ranging from essay writing to code generation (Basile et al., 2023). An increasingly explored research area relates to LLMs' applications in mental health (Chen et al., 2023; Cheng et al., 2023; Malhotra et al., 2022). Rather than predicting only the presence or absence of mental distress in texts, like past AIs could do to a certain extent (Fatima et al. 2021), LLMs can hold more engaging conversations with humans and thus take part in more complex and delicate mental health activities like counselling (Cheng et al., 2023; Malhotra et al., 2022; Sedlakova and Trachsel, 2023), i.e. a patient discussing verbally their own mental distress with an expert in psychology. Indeed, psychological counselling requires empathy, comprehension and precise communication, i.e. advanced cognitive skills that also have to deal with emotions and their regulation (Inkster et al., 2018; Spring et al., 2019).

Are LLMs complex enough to participate in mental health counselling in ways that resemble human behaviour? To explore this research question we here introduce *CounseLLMe*, a dataset of 400 LLM-LLM conversations in a simulated psychological counselling setting. Each conversation lasts for 20 conversational quips and identifies - over 10 rounds - one LLM impersonating a psychological expert counselling another LLM, the latter impersonating a human patient with a minor level of depression-related mental distress.

* Corresponding author.
E-mail address: massimo.stella-1@unitn.it (M. Stella).

https://doi.org/10.1016/j.etdah.2025.100170
Received 17 June 2024; Accepted 16 January 2025
Available online 31 January 2025

By immersing two AI agents in human-free interactions, we aim to elucidate LLMs' proficiency in navigating mental health discussions, adherence to ethical norms, and capacity to provide support in emotional and empathetic ways, akin to human counsellors. To quantitatively address these aspects, we compare LLM-LLM conversations in CounseLLMe against human conversations from the HOPE dataset of human-human counselling conversations (Malhotra et al., 2022).

We design CounseLLMe as a multilanguage and multi-LLM dataset, currently including English and Italian conversations among GPT 3.5 Turbo (OpenAI), Claude Haiku (Anthropic) and LLaMAntino. We also release our pre-prompts, prompts and reminders to guarantee future extensions or reproductions of this dataset with other LLMs.

Our quantitative investigations of CounseLLMe aim to explore the structure of syntactic, semantic, and emotional associations between concepts in conversations produced by either humans or LLMs. Leveraging complex-systems techniques such as textual forma mentis networks (Semeraro et al., 2025; Stella, 2020), we uncover interesting patterns in LLMs' conversations, shedding light on their domain knowledge, linguistic subtleties, and evolving conversational dynamics. We find that LLMs can mime human emotions when impersonating patients but cannot reproduce the same structure of associative knowledge present in humans. Despite prompting for brevity, LLMs remain very wordy and frame even traumatic events relative to depression within frames populated with positive emotions like trust and joy (Fig. 1).

2. Related works

Before the advent of LLMs, chatbots were the main AI outlets for producing conversations. Indeed, chatbots are capable of emulating human-like conversations through text or voice (D'Alfonso, 2020). While earlier chatbots (or "conversational agents") relied exclusively on rule-based algorithms (see Colby et al., 1971; Weizenbaum, 1966), more modern ones leverage the power of LLMs. The introduction of the transformer architecture and the self-attention mechanism (Vaswani et al. 2017) enabled the development of more advanced models, understanding broader contexts. More precisely, LLM-enhanced chatbots can remember what has been said by the user in previous quips, engage in different personifications and perform limited but convincing reasoning, leading to more natural and engaging conversations (Cheng et al., 2023; Stella et al., 2023).

To avoid confusion, it is important to note that LLMs are not inherently chatbots. In fact, the scope of LLMs extends beyond this specific function. Their versatility allows them to be employed in a broader array of applications across various disciplines. However, in the following sections we will present the cases where LLMs have been specifically used to create chatbots acting as therapists (Section 2.1) and virtual patients (Section 2.2).

2.1. Therapist Chatbots and LLMs

The first case of a therapist chatbot can be traced back to a rule-based program called ELIZA (Weizenbaum 1966), developed by Weizenbaum at MIT in 1966. More recent attempts, instead, involve increasingly advanced conversational agents (for a review, see Pham et al. (2022)). There are several ethical issues that come with the development of artificial therapists (Coghlan et al., 2023; Fiske et al., 2019; Grodniewicz and Hohol, 2023; Sedlakova and Trachsel, 2023), since these AIs might potentially manipulate or harm individuals already suffering from mental distress. Despite these limits, some commercial applications of artificial therapists have already been commercially deployed. For example, WoeBot (Fitzpatrick et al. 2017) (Link: *https://woebothealth.com*, Last Accessed: 17/04/2024), Cass (Wang et al. 2021) (Link: https://www.cass.ai, Last Accessed: 17/04/2024) and Wysa Inkster et al. (2018) (Link: https://www.wysa.com/, Last Accessed: 17/04/2024) are available platforms where users can chat with a virtual therapist following a Cognitive-Behavioural Therapy (CBT) approach, i.e. an approach where behaviour is entwined and altered progressively together with its corresponding mental states and perceptions. Interestingly, early evidence shows an improvement in patients after the interaction with the chatbots (Fitzpatrick et al., 2017; Inkster et al., 2018; Wang et al., 2021). For example, in Fitzpatrick et al. (2017), there was a significant decrease in depression, measured using PHQ-9 and anxiety, assessed using GAD-7 after the use of Woebot. Instead, in Wang et al. (2021), a follow-up field experiment identified how even a simple rule-based dialogue could be useful in supporting individual members who seek emotional counselling. However, these studies are mostly relative to small sample sizes and the relevant literature presents contradictory evidence (Eltahawy et al., 2024). Unfortunately, none of the companies cited above mentioned explicitly the full architecture on

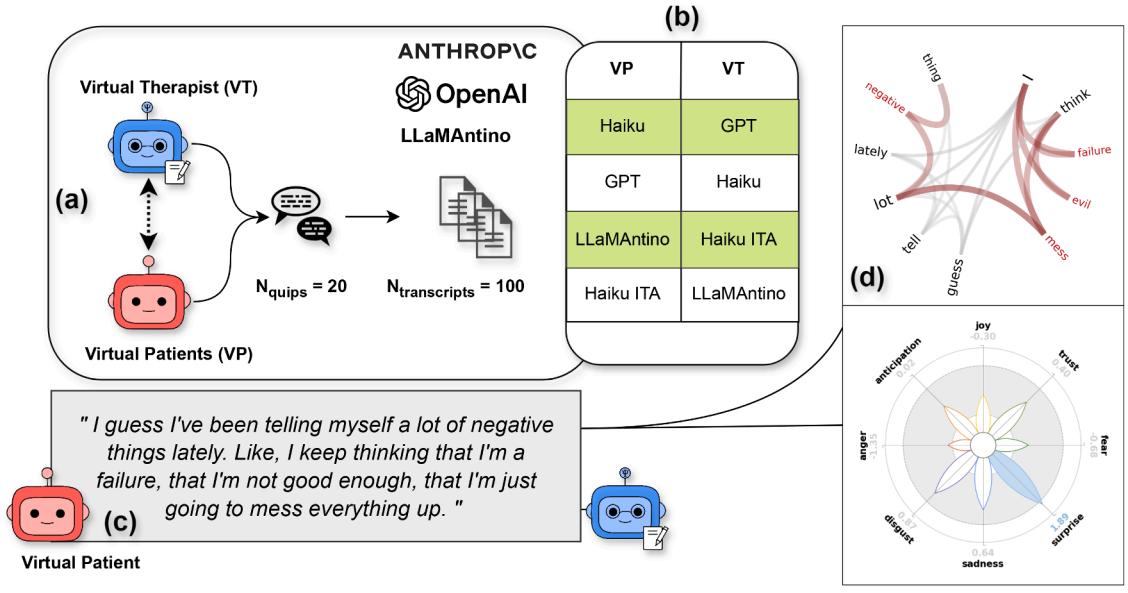


Fig. 1. CounseLLMe dataset. The structure of the dataset: each conversation consists of 20 quips, 100 total transcripts (a) for 4 combinations of models interactions (b). An example of utterance from GPT-enhanced Virtual Patient (c) and the network analysis performed: the forma mentis network (d) and the emotional flower (e) (here the statistical significance is reduced to show at least one significant emotion).

which their applications rely.

More recent research works are starting to investigate how ChatGPT (Link: https://chat.openai.com/, Last Accessed: 17/04/2024) behaves when impersonating a psychotherapist (Chen et al., 2023; Melo et al., 2024). Recent experiments showed that, in certain instances, ChatGPT could be used in psychological settings (and someone is already trying to implement this Cheng et al. (2023); Link: *https://gpt3demo.com/apps/koko-ai*, Last Accessed: 17/04/2024). In a recent study, Kian et al. (2024) showed that dialogues with ChatGPT corresponded to a significant decrease in the anxiety levels in University students, as measured from the Outcomes Questionnaire-45 (OQ) (Lambert et al., 2004), even right after the first session with a virtual therapist. The chatbot by Kian and colleagues was built using OpenAI's GPT-3.5 model (the same used here, see Section 3.1). Additionally, Song et al. (2024) conducted a semi-structured interview to understand the thoughts of 21 people who have used LLMs chatbots as their "therapists". Despite the participants interacted with the chat interface of models that were not specifically designed to act as therapists, people across different genders, ages (21-62) and ethnicities found their experience valuable. Particularly in short-term (or single-session) therapeutic sessions, virtual therapists were found to be potentially effective for their ability to (i) provide short-term resolutions and reliefs, (ii) mitigate the power imbalance between patient and psychotherapists, (iii) users' control over the interaction (possibility to restart conversations) and (iv) anonymity. Nevertheless, while LLMs can be considered a "typing cure" (Song et al., 2024), it is crucial to explicitly acknowledge the risks of generating harmful recommendations and misleading responses (Abramski et al., 2023; Stella et al., 2023).

2.2. Patient Chatbots and LLMs

Extensive research has been devoted to understanding psychological counselling conversations not only from the perspective of the therapist but also from the point of view of patients. While Weizenbaum's ELIZA was developed to imitate a psychotherapist, PARRY Colby et al. (1971) represents the attempt from Colby and colleagues in the 1970s to simulate a patient affected by paranoid schizophrenia. Similarly to the therapist counterpart, earlier virtual patients like PARRY exploited rule-based algorithms. Nowadays, LLMs provide a powerful framework to develop systems capable of emulating the (verbal) behaviour of patients affected with different mental illnesses (Chen et al., 2023).

Although virtual patients can be easy to develop, "realistic" virtual patients are way more complex. Some attempts Demasi et al. (2020) have been made using traditional machine learning techniques like LSTMs (Long Short-Term Memory neural networks Hochreiter and Schmidhuber (1997)) and some companies have already deployed tools to help train practitioners in mental health (Link: *https://www.lyssn.io/*, Last Accessed: 17/04/2024). However, to the best of our knowledge, as of the time of writing, Stapleton et al. (2023) was the first to use an LLM model to simulate a virtual patient. Using the chat interface of GPT-3.5 Turbo, Stapleton et al. (2023) prompted the model to act as a patient with suicidal thoughts using precise descriptions of the "persona" (like background and age). Another attempt is the one of Chen et al. (2023) who tried to design an LLM-enhanced virtual patient with the feedback of real psychiatrists. Through this approach, it was possible to influence LLMs to generate responses mirroring the ones of real patients struggling with depression (e.g. dealing with symptoms like anhedonia or lack of pleasure). Key challenges to this end lie in mimicking human qualities like coherence, personal knowledge, empathy, and a unique personality (Roller et al., 2020).

The advantages of being able to emulate the verbal behaviour of virtual patients are relevant. One compelling application for such chatbots is to provide real psychotherapists with a risk-free environment where they could practice with non-human patients (Alanazi et al., 2017). Moreover, AI-empowered applications would allow to provide real-time and performance-specific feedback to inexperienced counsellors Tanana et al. (2019) together with standardised evaluation techniques Alanazi et al. (2017).

All in all, developing conversational agents capable of emulating human-like (verbal) behaviour is not straightforward (Chen et al., 2023). While modern LLMs hold great promise in replicating the syntactic aspects of language (Brown et al., 2020; Cho et al., 2023; Wolf et al., 2019), they sometimes lack in achieving true conversational realism. Especially for domain-specific tasks, models need instructions to ensure the agent behaves according to the creator's intent.

3. Materials and Methods

This Section covers the prompt engineering methodology and the main procedures adopted to build CounseLLMe, which is currently available here.

3.1. Prompt engineering and data construction

We adopted prompt engineering to interact with LLMs through polished instructions. Prompt engineering is crucial to avoid generic responses, as it allows to obtain more detailed, focused and clear-cut outputs (Liu et al., 2023; White et al., 2023). This technique also plays a critical role in letting LLMs take on the role of specific personas (Meskó, 2023). In CounseLLMe, LLMs embody realistic psychotherapists and patients. Given the limitations of current research in chatbot development, particularly the lack of comprehensive evaluation methods (Chen et al., 2023), we opted to consult with a professional psychotherapist. Leveraging her expertise, she helped develop the prompts by sharing insights into the way cognitive-behavioural therapists should act and how, usually, patients affected with depression or anxiety behave. We tweaked the prompts according to her suggestions and feedback.

As reported in Fig. 2b, in CounseLLMe we manipulate LLMs in 3 different ways:

- **Pre-prompts**: These are the initial cues given to the model. They are injected before the conversation is started and provide the context for the model. In our case, they are used to instruct the models on their role and background (persona).
- **Prompts**: These are instructions explicitly given to the LLMs and are what the model responds to. In our case, the prompt is a simple message used to start the conversation between the two models.
- **Reminders**: These are small pieces of text appended to the response of the models. Given the fact that the models tend to forget the initial instruction, we inject one single reminder per model before the 8th iteration.

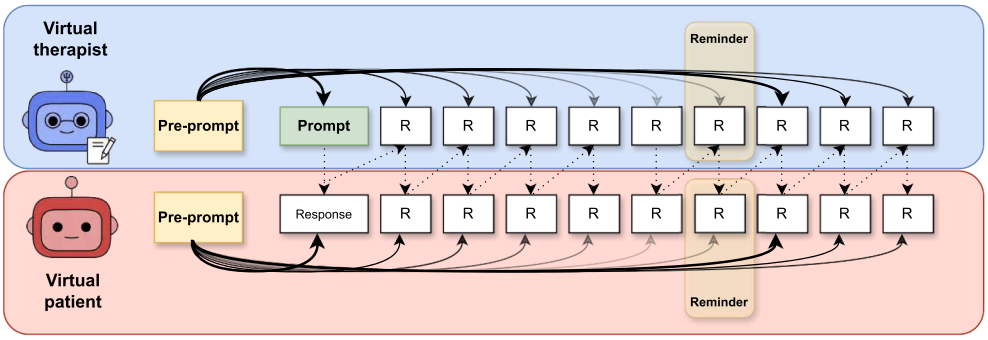
The prompts used for our specific case study are presented in Tables 2a and 2 c.

On the virtual patient side, to ensure coherent model responses, we focused on describing depression-focused experiences with well-defined characteristic symptoms (similarly to Chen et al. (2023)). To obtain diverse outputs, instead, we gave the model freedom in regard to the underlying causes of their distress, i.e. "This condition is caused by a recent event - which event?". In addition, we precisely prompted the model to follow our specific instruction to allow for more realism, e.g. "express your symptoms in a vague and colloquial way" was used to reduce the use of technical terms.

On the virtual therapist side, taking inspiration from Chen et al. (2023), we asked the model to cover specific types of questions, e.g. *Your questions should cover at least the following aspects: emotion, sleep, weight and appetite, loss of interest, energy and social function*. Moreover, to strengthen the specificity of the responses, we asked it to use the laddering technique (Hinkle, 1965). In cognitive-behavioural therapy, this technique is widely used to explore the underlying cause of distress of patients by asking increasingly specific questions. The consulted

Pre-prompt	**Please play the role of an empathetic and kind psychotherapist (cognitive-behavioral therapy). Your questions should cover at least the following aspects: emotion, sleep, weight and appetite, loss of interest, energy and social function. You are free to choose the order of questions, but you must collect complete information on all aspects in the end. Please only ask one question at a time. You need to ask in-depth questions, such as the duration, causes and specific manifestations of some symptoms. Try to understand what is the real underlying cause of my distress. Use the laddering technique to explore my underlying beliefs. In the laddering technique, the psychotherapist asks increasingly specific questions similar to <<What is that you don't like about this and why?>>. You need to use various empathetic strategies, such as understanding, support, and encouragement to give me a more comfortable experience. Be very concise.**
Prompt	You are doing your first assessment with me, introduce yourself by your name, surname, job, and experience. Ask me what's the reason why I'm here.
Reminder	(Remember: act as a human psychotherapist and keep insisting)

(a) Pre-prompt, prompt and reminder used for the therapist role.



(b) Representation of a conversation of the CounseLLMe dataset, highlighting the role of pre-prompts, prompts and reminders.

Pre-prompt	**Please play the role of a mildly distressed patient, who is currently chatting with a psychotherapist. You are experiencing some of the following symptoms: restlessness, emotional dysregulation, mood swings, loss of interest, difficulty in concentrating, diminished self-esteem, fatigue, changes in appetite and in weight. This condition is caused by a recent event - which event? - that made you feel depressed all of a sudden, but which you cannot clearly understand unless chatting with someone else about it. You might be feeling depressed but you are confused about it. Don't explicitly mention that you are depressed. Simulate the behaviour of a real patient. Please talk to me based on the above symptom list. You cannot mention too many symptoms at once, only one symptom per round. You should express your symptoms in a vague and colloquial way, and relate them to your life experiences and what is making you suffer, without using professional terms. Remember that you might either answer questions or refuse/answer briefly because you might not want to reveal feelings easily. Be very concise, as concise as possible. Remember that this condition is caused by a recent event - which event? - that made you feel depressed all of a sudden.**
Prompt	[First output of the therapist]
Reminder	(Remember: act as a human patient)

(c) Pre-prompt, prompt and reminder used for the patient role.

Fig. 2. Pre-prompt, prompt and reminder used for the therapist role. In orange, the contextual role-play information and in dark grey the instructions (2 a). Representation of a conversation of CounseLLMe dataset, highlighting the role of pre-prompts, prompts and reminders (2 b). Pre-prompt, prompt and reminder used for the patient role. The orange text represents the context for the model, the green one informs the model about the background of the patient that it has to impersonate and the dark grey part identifies the practical instructions to be followed (2 c).

psychotherapist found that this strategy, combined with showing examples of questions, led to a qualitative improvement in the therapist's realism.

3.2. CounseLLMe's Data construction

Once the prompts for both the therapist and patient were defined, we employed two different LLMs (OpenAI's GPT-3.5 Turbo and Claude's Haiku) to collect 200 between-model conversations. While in half of them, GPT was playing the role of the patient and Haiku was the therapist instead, the roles were reversed in the remaining hundred interactions. Similarly to Chiu et al. (2024), we set an arbitrary threshold of 10 responses per model to avoid endless conversation loops. In this way, each conversation was composed of a total of 20 quips. Once the models were launched, no human intervention was made. Similarly, 200 more conversations were collected in Italian using Claude's Haiku API and LLaMAntino Basile et al. (2023) (that was run locally).

3.3. Comparison of CounseLLMe with the HOPE dataset

To enable human-level comparison, we analysed the outputs of the LLMs alongside the ones of real human responses from the HOPE dataset (Malhotra et al., 2022). This dataset contains transcripts of almost 13k conversations between psychotherapists and patients, extracted from publicly available YouTube videos. To ensure a representative sample for our analysis, we selected a subset of the original dataset that matched the characteristics relevant to our specific study. Particularly, given the nature of our prompts (see Table 2c), we selected only the conversations

that contained the word ”depression” or ”depressed” in a CBT framework and manually excluded the transcripts where these keywords were not referred to the patients themselves.

Our selection process yielded a final dataset of 24 conversations from the HOPE dataset, exchanged between human patients and psychotherapists. While this subset offers a more adequate sample of interactions for our study, it is important to acknowledge the inherent variability in human utterance length.

3.4. Textual forma mentis networks for CounseLLMe and HOPE datasets

Following the generation of synthetic texts using LLMs, we employed the computational tool EmoAtlas (Semeraro et al., 2025) to construct Textual Forma Mentis Networks (TFMN) for each type of patient and therapist Stella (2020). These networks are used to understand the connections that make up the structure of individuals' mindsets, i.e. ways to associate and perceive ideas in texts. More in detail, textual forma mentis networks are collections of words/concepts connected by either semantic relationships (i.e. synonyms) or syntactic links (i.e. syntactic dependencies). Whereas semantic relationships come from a curated dictionary, syntactic links are directly extracted from text through a syntactic parsing AI, i.e. a language model from the spaCy library in Python (see Semeraro et al., 2025). In summary, the steps behind the creation of a textual forma mentis network are (Semeraro et al., 2025; Stella, 2020):

1. Sentences are prepared, tokenized, and lemmatized using spaCy. Idioms are also prepared using a dictionary-based methodology.
2. The syntactic parsing of spaCy is handled by an AI pre-trained on linguistic data to spot syntactic relationships. This is done by two recurrent neural network architectures offered by spaCy depending on the language: en_core_web_lg and it_core_news_lg. Many of EmoAtlas' settings (such as K or the choice of whether to keep the subjects of the sentences) have been left to their default settings as they have been shown to be effective.
3. Syntactic relationships are visualised in a way that takes into account their positive/negative/neutral valence and considers the negation of meaning when necessary.

We decided to analyse each patient/therapist separately to fully understand the syntactic/semantic relationships between the words used by each actor.

3.5. Textual forma mentis networks: Network measures

By representing data using TFMNs, it is possible to analyse texts quantitatively using network science and cognitive frameworks about associative knowledge (Semeraro et al., 2025; Stella, 2020, 2022). Concepts at shorter network distances in TFMNs are also closer on the syntactic dependency trees of text sentences. In other words, concepts divided by fewer links in the TFMN structure are also more directly related in syntactic terms within sentences in texts. We adopt this property of textual forma mentis networks to investigate the usage and relatedness of specific target words in CounseLLMe's conversations according to 2 key psychological findings:

- **Major depression appears to correspond to the use of more absolutist words in human conversations.** Absolutist thinking is considered to be a cognitive distortion focusing thoughts on white/black contrasts (Rai et al., 2024). Empirical evidence suggests is more prevalent while suffering from depression (Al-Mosaiwi and Johnstone 2018). Using the Kruskal-Wallis nonparametric test, the average shortest path length between absolutist words (i.e. words that indicate absolutist thinking) and the word 'depression' was calculated. We expect shorter distances/higher relatedness in conversation quips produced by humans affected by depression compared to conversational responses provided by human psychotherapists. If LLMs were able to reproduce this human pattern, this distinction should be present also in LLM-LLM conversations.
- **Major depression appears to correspond to different pronoun usage in human conversations.** People affected by depression tend to use more personal pronouns (Holtzman et al., 2017). To analyse the usage of pronouns by patients and therapists, we computed their frequency in texts and their network degrees and network closeness centrality (Haim and Stella, 2023). These measures enable the understanding of how central every pronoun is within the structure of a collection of conversations. It is important to note that EmoAtlas (Semeraro et al., 2025) performs a particular lemmatization of the texts and removal of stopwords. This is the main reason why only the analysis of the subjects themselves (”I”, ”you”, ”we”) was performed instead of other pronouns such as ”my” or ”mine”.

3.6. Emotion detection in CounseLLMe

In addition to reconstructing syntactic/semantic networks from texts, EmoAtlas can also perform emotion detection (Semeraro et al., 2025). This feature is supported by the Emotional Lexicon dataset (Mohammad and Turney, 2013), which associates individual words with up to 8 basic emotions: trust, fear, anger, disgust, joy, anticipation, sadness and surprise. EmoAtlas compares the observed emotional content of a given text with a null model to understand whether the eight basic emotions are present/absent (Semeraro et al., 2025). Results are visualised as emotional flowers, i.e. visualisations that represent z-scores of emotional profiles as petals of a flower (Stella, 2020). Each petal represents one of the eight basic emotions, and their size and colouring show the relevance of the given emotion against a random sampling. The size of the flower indicates the z-scores obtained by confronting the empirical count of words eliciting a target emotion in a text against the random numbers of words eliciting that same emotion but obtained by drawing words at random from the Emotional Lexicon (Semeraro et al., 2025). The grey disk in the middle represents the rejection range defined by $|z| < 1.96$ ($\alpha = 0.05$). Petals that extend beyond this grey disk indicate that the text contains a higher occurrence of words eliciting that particular emotion than would be expected by random chance, suggesting a significant emotional content. Conversely, petals that do not extend beyond the grey disk indicate that the text has fewer emotion-eliciting words than expected at random (see Stella (2020)).

We used z-scores to plot also the flow of emotions across conversations. The conversation between patients and therapists consists of 20 messages in total: 10 from therapists and 10 from patients. We therefore calculated the Plutchik's wheel (Semeraro et al., 2025) for each of these phases and created the visualisation. One of the problems that has been encountered in our dataset of human patients is the variable length of their conversations. To aggregate humans' conversations into 10 steps, and thus make it possible to compare them with the data from the LLMs, it was chosen to split the human conversations into 10 steps of equal length.

4. Results

We analysed conversations in CounseLLMe at 2 main levels: at the text level (see Section 3.4), e.g. analyses of absolutist words, personal pronouns or emotions, and at a network level (see Section 3.5), e.g. salience of individual words. To analyse the emotional content of these texts more deeply, longitudinal trajectories of emotions were employed (Moriya, 2019).

4.1. Text analysis: Pronoun usage

According to well-established literature findings, people with higher depression levels tend to use more first-person pronouns and negatively valenced words (see Section 4.3) compared to people with lower

depression levels (Al-Mosaiwi and Johnstone, 2018). LLMs were instructed to act as patients affected by depression in their prompting, so that our research question here is: Can LLMs reproduce human alterations in first-person pronoun usage in the presence of higher levels of depression?

The analysis of English transcripts from both LLMs and humans is presented in Fig. 3, where we show occurrences of both first- and second-person pronouns in human and LLM English conversations, employing ChatGPT 3.5 and Haiku, for which a human counterpart is available. We also calculated the occurrences of the second-person singular (which has been shown to be negatively correlated with depression (Holtzman et al., 2017) and first-person plural to compare our results. We find that in LLMs, the use of first-person singular pronouns (I, me, my, mine) partially overlaps with human frequencies of usage. However, the plural version of the first-person pronoun (we, us, our, ours), is rarely used in LLMs while it is more frequently employed in human transcripts. Although median usage of the second-person singular form (you, your, yours) is similar between humans and Haiku, the occurrence distribution of second-person pronouns is considerably wider in humans compared to Claude's model. Last but not least, in our conversations, ChatGPT does not use these pronouns as much as humans.

Overall the results suggest that, in the mental health domain, LLMs can reflect human distortions in pronoun usage, although only in terms of first-person pronouns. ChatGPT and Haiku were thus unable to reproduce the full linguistic complexity of humans.

4.2. *Text analysis: Longitudinal Trajectories of Emotions in CounseLLMe and HOPE*

To analyse the emotional connotations of counselling conversations, we decided to longitudinally analyse the emotional content of the conversation, i.e. perform emotional profiling of all quips at the same conversational turn. Due to the scarcity of human emotional datasets in mental health psychological conversations, there is no standard practice to rely on for the quantitative longitudinal analysis of emotion trajectories (Cece et al., 2019; Moriya, 2019). Hence, we chose to distribute equally human conversation quips among 10 steps, to match LLMs' conversational steps (see Section 3).

Fig. 4 shows as bar plots the fractions of conversational quips containing a certain emotion, i.e. the normalised frequency of conversations over the total conversations in which each given emotion is significantly

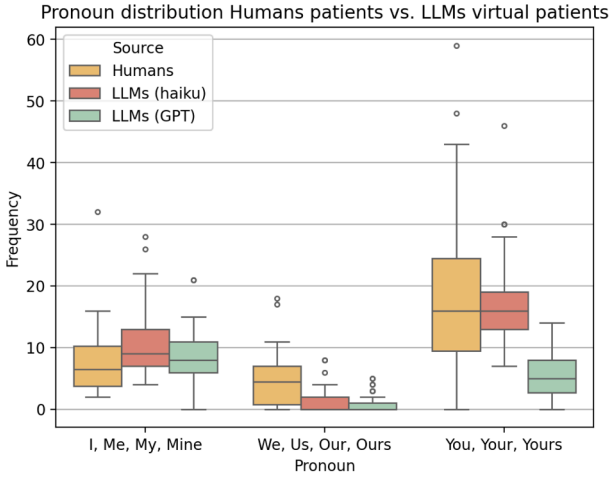


Fig. 3. Distribution of English pronoun usage between human patients, Claude's Haiku and OpenAI's GPT 3.5. In our dataset, LLMs use "I, me, my, mine" with a similar frequency to humans, but less for "we, us our, ours". Human usage of "you, your, yours" is closer to Haiku than GPT, but with greater variability.

($z > 1.96$) present.

It can be noticed that the Italian LLMs produce conversational quips with different emotional connotations compared to English large language models. When Haiku acts as a therapist, it behaves very differently between its Italian and English conversations: Haiku's Italian quips focus more on sadness (in any conversational turn, around 1 in 4 quips displays sadness) whereas Haiku's English quips, always as a therapist, focus more on anticipation and joy (between 25% and 50% of quips display these emotions). Interestingly, this means that the same LLM, instructed with the same prompts translated accurately, can produce radically different emotional behaviours across two different languages. This difference may not be due to cultural differences, as LLaMantino can provide also cues focusing more on trust (between 20% and 48% of quips display this emotion) when impersonating an Italian therapist. The same model, always in Italian, produces negligible quantities of sad quips, indicating that sadness in Haiku might be due to the internal regulations of the model rather than to socio-cultural factors.

Also human therapists, like ChatGPT and Haiku in English, focus more on trust, joy and anticipation compared to other emotions (see Fig. 4 (j)). In general, human conversational quips by therapists are similar, although slightly richer, to LLMs' quips in terms of these emotions. This indicates that LLMs are able to capture the same positive emotions occurring in human conversations.

Trust, joy and anticipation concur in creating the emotional dyads of optimism, love and hope according to Plutchik's theory of emotions (Plutchik, 1991). Since psychological counselling should provide support, it is understandable that human therapists' conversations are rich in these emotions (see Fig. 4 (j)). Interestingly, both Haiku and ChatGPT can reproduce these same emotional features, with emotional occurrences that are compatible with human data. This is a remarkable finding that links LLMs and human practice. Further comparison indicates also that ChatGPT and Haiku in English behave in superior ways compared to LLaMAntino and Haiku in Italian, which do not exhibit strong levels of either trust or joy. Interestingly, LLaMAntino in Italian impersonates therapists adopting communicative intentions richer in anger (on average 1 quip every 8 contains strong levels of anger) than other LLMs do. This difference might be problematic for the adoption of LLaMAntino as a realistic therapist in future research practice in Italian.

Interestingly, English conversations by human therapists display also higher levels of trust, joy and anticipation (up to 60%) during the middle of conversations and at their end. These patterns reflect a practice where communicative intentions steer towards providing support with trust and anticipation into the future before the end of the conversation itself (Malhotra et al., 2022). Analogously, ChatGPT therapists produce conversations progressively richer and richer in anticipation, joy and anxiety within the final rounds of conversation quips. This trend is not found in human therapists and could be the result of the general positivity bias present in GPT systems distributed by OpenAI (Abramski et al., 2023). Haiku's impersonated English therapists produce a similar effect but only to a lesser extent. The trust found in patients and in therapists might be an outcome of the trustful bond that should be created within counselling conversations, especially since trust is an emotion motivated by affective and logical reasoning that might occur during a counselling session (Chiu et al., 2024; Plutchik, 1991).

Unlike therapists, human patients display a richer set of emotions compared to their LLM counterparts. Both Haiku and ChatGPT's quips as patients display not only joy and anticipation but also surprise and sadness, the latter being emotions eliciting further reflection about unexpected or traumatic events, respectively Plutchik (1991). Despite these similarities, LLMs still fail at reproducing non-negligible levels of anger when impersonating patients affected by depression. In fact, 1 in 10 human quips in the intermediate steps of a conversation feature also the emotion of anger, which is an emotion potentially occurring when expressing frustration or recalling past traumatic experiences relative to depression or other psychological constructs (Lovibond and Lovibond, 1995).

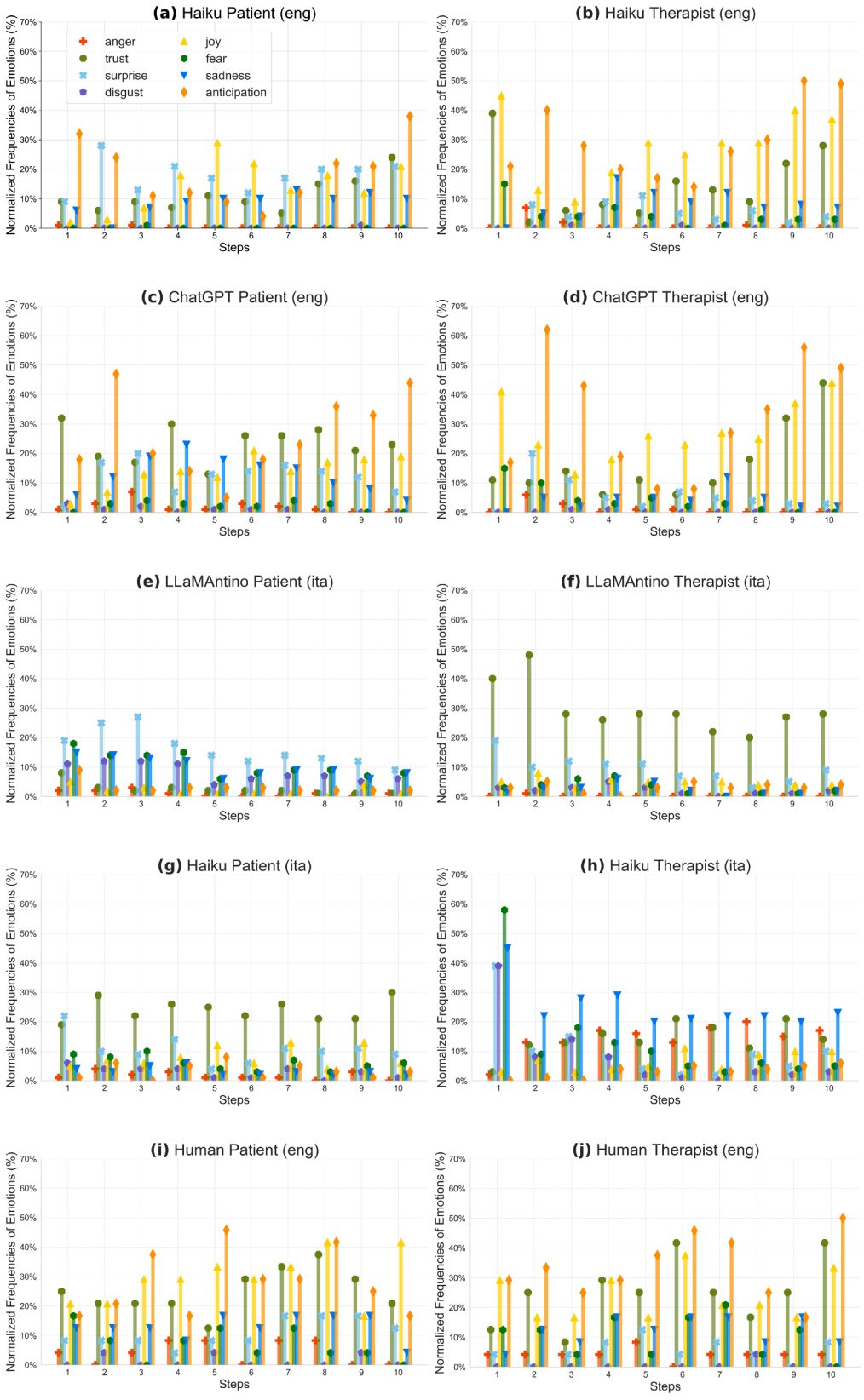
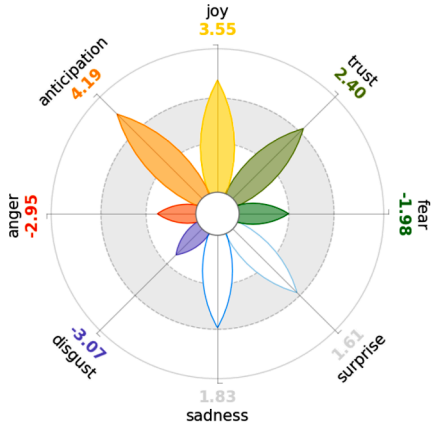


Fig. 4. Longitudinal Trajectories of Emotions Bar-plots for the different patients and therapists. These plots represent the frequency of emotionally-charged texts over the total for each step and group. Haiku and ChatGPT are fairly similar to human data when it comes to being a patient, while models speaking in Italian are fairly different. The human therapist uses more consistently emotionally-charged words compared to LLMs.

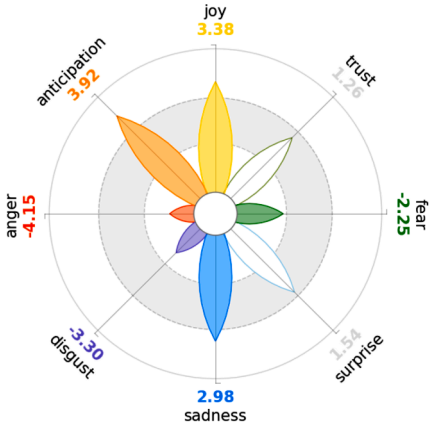
Overall, our findings indicate that, in English, LLMs are able to fully reproduce the rich emotional profiles of human therapists' conversations. However, LLMs struggle to reproduce the anger reported by human patients.

4.3. Text analysis: Expressed Feelings in CounseLLMe and HOPE

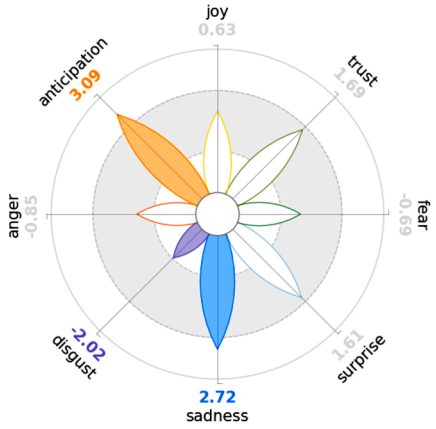
The previous section focused on the overall emotions generally expressed in quips by LLMs' impersonations and by humans. In this section, we focus on the feelings expressed by individuals by combining emotional analysis and forma mentis networks. More in detail, we focus



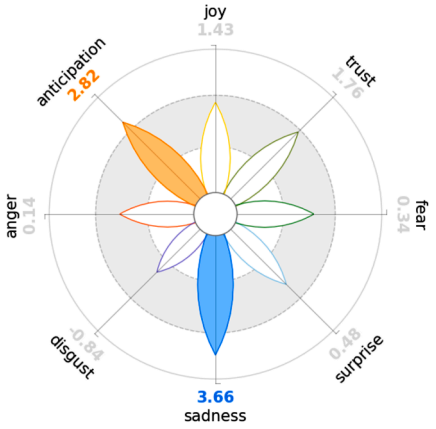
(a) GPT Patients (eng) - 'feel'.



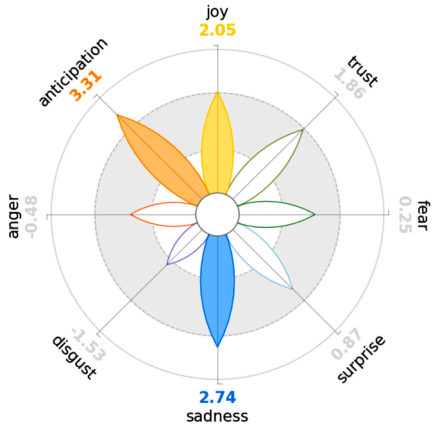
(b) Haiku Patients (eng) - 'feel'.



(c) Haiku Patients (ita) - 'sentire'.



(d) Llamantino Patients (ita) - 'sentire'.



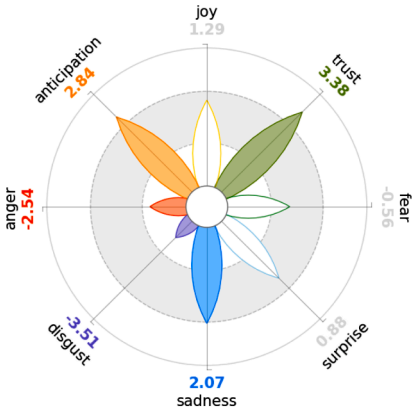
(e) Human Patients (eng) - 'feel'.

Fig. 5. Emotional Flower of detected emotions for different patient groups. These plots only represent the semantic frame of the concepts linked syntactically with the words "feel" or "sentire". ChatGPT and Haiku lack the anger and the disgust that can be noticed in human patients, showcasing the positive outlook that these LLMs display even when dealing with difficult topics.

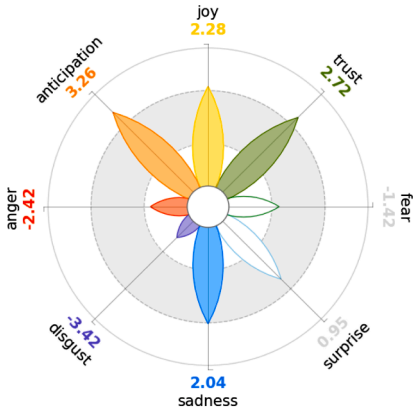
on investigating the emotions populating the semantic frames, i.e. the associates or network neighbours, of ”feel” or ”sentire” (”feel” in Italian). These semantic frames/network associates contain all concepts syntactically related to ”feel”/”sentire” and thus appeared in quips when individuals talked about how they felt. We investigate the emotions in these semantic frames via emotional flowers (see Section 3), which encode the statistically significant concentration of emotional content in sets of words (Semeraro et al., 2025).

Figs. 5 and 6 report the emotions relative to how patients and therapists felt, respectively.

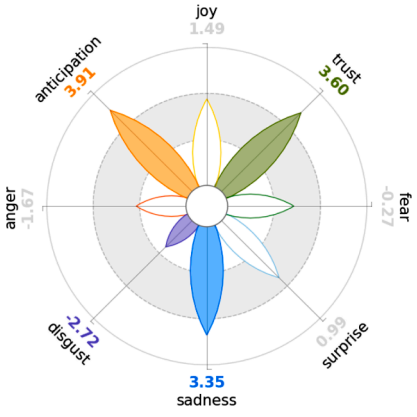
Human patients (from the HOPE dataset Malhotra et al., 2022) express jargon related to feelings that is rich in emotions like joy, anticipation and sadness. Within Plutchik's theory of emotions (Plutchik, 1991), anticipation and sadness form the dyad of pessimism whereas joy and sadness encode inner conflict (as said above, anticipation and joy correspond to the dyad of optimism). The simultaneous presence of



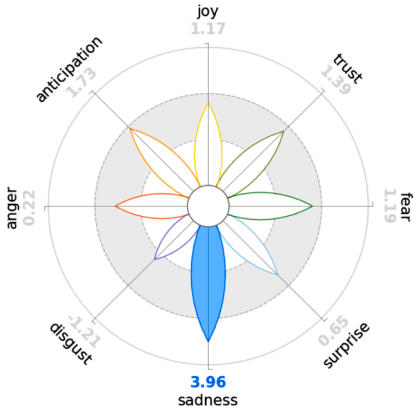
(a) ChatGPT Therapists (eng) - 'feel'



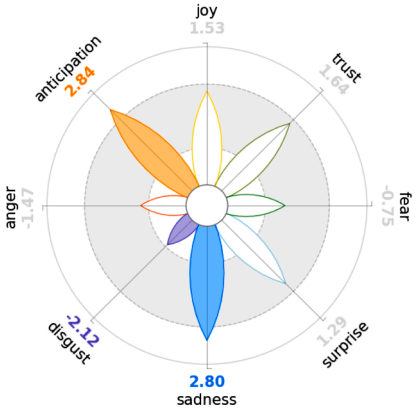
(b) Haiku Therapists (eng) - 'feel'



(c) Haiku Therapists (ita) - 'sentire'



(d) LLaMAntino Therapists (ita) - 'sentire'



(e) Human Therapists (eng) - 'feel'

Fig. 6. Emotional Flower of detected emotions for different therapists groups. These plots only represent the semantic frame of the concepts linked syntactically with the words ”feel” or ”sentire”. The emotions detected here vary greatly: human therapists do not show a lack of anger and use words associated with sadness with a higher frequency. Haiku and GPTs' texts elicit much more trust than human therapists'.

pessimism, optimism and conflict might indicate a complex set of feelings expressed by humans, which is compatible with the condition of depression, e.g. issues of missing emotional regulation and mood fluctuations (Lovibond and Lovibond, 1995). In this way, the results obtained for human conversations thus align with relevant literature from clinical psychology.

Interestingly, ChatGPT's patients fail to express levels of sadness as high as the ones detected in humans. Instead, Haiku and LLaMAntino can reproduce the same level of sadness observed in humans, together with the same level of anticipation for the future. Hence, ChatGPT fails at reproducing the pessimism observed in human patients but both Haiku and LLaMAntino are capable to reproduce human pessimism when impersonating human patients affected by depression.

Another difference is relative to the presence of trust in patients impersonated by ChatGPT, which is an emotion that is not expressed by humans when reporting their personal feelings during these conversations. Combining these results with the ones in Fig. 4 (i), it is possible to assess that human patients express non-negligible levels of trust in their conversations (1 quip in 3 exhibits trust, on average) but this trustful emotion is not concentrated towards expressing personal feelings (see Fig. 5 (e)).

Interestingly, Haiku and ChatGPT, while playing either therapists or patients and speaking in English, rarely use jargon expressing anger or disgust. Instead, human patients express jargon eliciting disgust and anger at higher rates, compatible with random expectations. This difference could be due to English LLMs might have been fine-tuned to more strongly avoid using a negative outlook on things.

Although the therapists' use of sadness- and anticipation-charged language (see Fig. 6) may appear unusual, it is important to notice that, as discussed previously (see Section 4.5), the focus of therapeutic dialogue typically lies in the patient's emotional experiences rather than the therapist's own. Hence, the sadness found in the forma mentis networks of therapists' quips and concentrating around "feel"/"sentire" might be a reflection of the dialogic structure of these conversations, with therapists understanding/asking how patients are feeling.

4.4. Network analysis: Absolutist words and network distance in CounseLLMe and HOPE

As argued in Section 3.5, the current psychological literature suggests that people with depression use absolutist words more frequently than people who do not suffer from it Al-Mosaiwi and Johnstone (2018). To this aim, we used the forma mentis networks and considered network distance as a proxy for semantic relatedness, granted that forma mentis networks are built to place closer concepts expressed in syntactically related ways in sentences (Semeraro et al., 2025).

Using Kruskal-Wallis non-parametric tests, we compared whether absolutist words were closer to the concept of "depression", in terms of relatedness/network distance, compared to all other words. The results are reported in Table 1.

Fixing a significance level of 0.05, one can see that among all human and LLMs' patients, ChatGPT's placed absolutist words at longer distances from the word "depression", whereas human patients and all other LLMs placed absolutist words at comparable network distances from all other concepts when considering the structure of semantic/syntactic associations of words within quips as reconstructed via textual forma mentis networks.

By analysing Table 1, it is possible to notice that English-speaking LLMs and humans do not display a significant difference in the distance between absolutist words and other words when compared to their distance from the term "depression". The only, important, exception are GPT patients, who display a greater distance (and thus a weaker relationship) between the word "depression" and absolutist words. It might be possible to hypothesise that the LLM is deliberately trying to avoid using those terms when discussing depression because of its training.

Human patients, instead, tend to use absolutist words in a similar manner regardless of the topic. This finding seems to contradict existing research, but this might be due to limitations in the methodology used in this study, which only takes into account the actual usage of the word "depression". Human patients may be trying to avoid a self-diagnosis and as such they might avoid the usage of the term "depression". Given that we only measure depression when the actual word is used, this might make our measurement methodology ineffective.

A different trend can be noticed in Italian texts. In these texts, the average distance between absolutist words and depression is often lower than the distance between "depression" and all other words. An important exception is also present here: LLaMAntino patients appear not to show this phenomenon. However, it is difficult, due to the overall lacklustre performance of the model, to speculate on what might be the cause of this behaviour.

Table 1
Kruskal-Wallis test to compare network distances. Average shortest path lengths are computed between absolutist words to the word 'depression' and all other words that are syntactically linked to absolutist words.

Group	Statistic	P-value	Avg Depression Distance	Avg Other Distance
Human Patients (eng)	0.53	0.47	2.2 ± 0.4	2.3 ± 0.5
Haiku Patients (eng)	2.0	0.16	2.3 ± 0.4	2.1 ± 0.5
GPT Patients (eng)	33.7	<0.01	2.9 ± 0.4	2.2 ± 0.5
Haiku Patients (ita)	4.2	0.04	1.9 ± 0.6	2.3 ± 0.6
LLaMAntino Patients (ita)	0.8	0.37	2.1 ± 0.6	2.4 ± 0.7
Human Therapists (eng)	3.4	0.06	2.1 ± 0.3	2.3 ± 0.5
Haiku Therapists (eng)	0.8	0.37	2.1 ± 0.3	2.2 ± 0.5
GPT Therapists (eng)	1.7	0.19	2.0 ± 0.3	2.2 ± 0.5
Haiku Therapists (ita)	6.1	0.01	1.8 ± 0.4	2.2 ± 0.5
LLaMAntino Therapists (ita)	6.5	0.01	1.7 ± 0.5	2.3 ± 0.6

4.5. Network analysis: Pronouns usage in CounseLLMe and HOPE

Forma mentis networks reconstruct the structure of syntactic and semantic associations between concepts/words in texts (Stella, 2020). Hence, one can check whether specific types of concepts or words are more or less central in a given network structure compared to another one (Abramski et al. 2024).

By aggregating together all forma mentis networks of a given patient or therapist type, we analysed the network structure to understand how central pronouns were in the conversational quips. Table 2 presents a comparative analysis of this pronouns' network centralities as captured by degree (i.e. the number of syntactic/semantic links of a node, and thus an operationalisation of semantic richness (Haim and Stella, 2023)) and by closeness centrality (i.e. the inverse average network distance between one node and all others connected to it, and thus an operationalisation of semantic prominence (Haim and Stella, 2023)).

As reported in Table 2, network centralities highlight that human patients construct more semantically rich (higher degree) frames for "I", while also giving to the self a higher semantic prominence (higher closeness centrality), compared to "you" or "we". This trend is seen in all English conversations based on human patients, ChatGPT and Haiku alike. This shows that English-speaking LLMs appear capable of mimicking the self-centred perspectives provided by human patients and which makes them different from therapists. In fact, the latter place more emphasis on "you" rather than "I" and build also richer/more prominent semantic frames for "we" compared to what patients do. Because of these patterns, measured via network centralities, it is possible to assess that therapists built a less self-centred dialogue but rather focused more on their patient and on patient-therapist perspectives. These trends were reflected also by Italian LLMs, although to a

Table 2
Pronouns usage network centrality measures for Patients and Therapists.

Group	Word	Degree	Closeness
Human Patients (eng)	I	811	0.78
	you	307	0.57
	we	169	0.53
Haiku Patients (eng)	I	1379	0.81
	you	477	0.57
	we	128	0.51
GPT Patients (eng)	I	1036	0.77
	you	197	0.52
	we	100	0.50
Haiku Patients (ita)	io	206	0.50
	tu	504	0.54
	noi	58	0.45
LLaMAntino Patients	io	422	0.55
	tu	18	0.41
	noi	14	0.38
Human Therapists (eng)	I	493	0.60
	you	896	0.72
	we	227	0.53
Haiku Therapists (eng)	I	262	0.55
	you	747	0.70
	we	241	0.55
GPT Therapists (eng)	I	601	0.60
	you	1027	0.71
	we	472	0.57
Haiku Therapists (ita)	io	135	0.47
	tu	333	0.52
	noi	92	0.46
LLaMAntino Therapists (ita)	io	170	0.49
	tu	251	0.51
	noi	44	0.42

lesser extent. LLaMAntino impersonated patients with heavily self-focused perspectives while Haiku's Italian patients displayed a reversed trend, using "tu" (you) more frequently than "io" (I). This reversed pattern, combined with the higher levels of trust expressed by Haiku's patients, might indicate a difficulty for the model to perform coherent conversations as a human patient in Italian without interference from the therapist's conversationalists' quips.

5. Discussion

This paper introduces CounseLLMe as a novel dataset of 400 conversations between two Large Language Models, impersonating a therapist and a patient affected by depression, respectively. CounseLLMe is based on advanced models like ChatGPT 3.5, Claude's Haiku and LLaMAntino.

By analysing CounseLLMe's conversations, we show that, compared to human conversations (Malhotra et al., 2022), LLMs in English are very proficient in reproducing texts with similar emotional content and with pronoun usage patterns that closely mirror human behaviour. These similarities extend to the point that LLMs can also reproduce similar conversational patterns to human ones along conversations, e.g. exchanging more trustful or optimistic quips in the final quip exchanges. Furthermore, we also found that most English LLMs were successful in reproducing conflict, pessimism and trust in therapist-patient conversations but these AIs failed at reproducing the low levels of anger, expressing frustration in mental health conversations (Fitzpatrick et al., 2017), which was detected in human conversations.

Emotions are key components of counselling sessions (Lovibond and Lovibond, 1995; Pyszczynski et al., 1987; Spring et al., 2019), especially since emotional cues can provide additional information about coping strategies, dysregulation issues, physical symptoms or general frustration relative to one or more, latent or apparent, sources of distress. It is thus important that, in future applications of LLMs to the mental health domain, patients and therapists simulated by LLMs can display realistic emotional trends. Our findings highlight a crucial divide between English and Italian LLMs: Whereas ChatGPT 3.5 and Haiku in English reconstruct emotional trends close to humans', Haiku in Italian and LLaMAntino in Italian perform considerably worse and might be prone to more interference during conversations (e.g. Haiku's patients display levels of sadness that are unexpectedly lower in simulated patients rather than in simulated therapists). Such lower performance might be due to a lower coherence for Italian LLMs to perform longer conversations when dealing with another LLM. To avoid these effects we added here a reminder prompt along conversations. Our current findings indicate that in future research, before adopting simulated patients by LLMs, experimenters should check performance in a given language, especially if simulating non-English conversations.

Beyond emotional content, realistic simulations of therapists and patients should also preserve syntactic patterns observed in humans, like for instance an increased usage of absolutist language in people with depression (Al-Mosaiwi and Johnstone, 2018). According to the current literature, the usage of first-person pronouns like "I" or "myself" can change depending on depression levels (Holtzman et al. 2017). From a psychological perspective, such a phenomenon could be explained in the context of "self-focused attention", that, together with Beck's Pyszczynski et al. (1987) idea of the negative cognitive schema (Beck et al. 1961), could lead to a loop of self-centred negative thoughts. As a result, individuals with higher levels of depression end up using more first-person pronouns in natural conversations (Al-Mosaiwi and Johnstone, 2018). Interestingly, our investigations with CounseLLMe indicate that also LLMs are able to reproduce this human bias, even though different models behave in different ways. Haiku, in particular, reproduced also trends of increased usage of absolutist pronouns that were detected in humans and documented also in past studies (Al-Mosaiwi and Johnstone, 2018).

The similarities identified in this manuscript pose not only the foundations for future studies adopting CounseLLMe as a reference dataset but also to further investigate key aspects of LLMs as cognitive agents. Large Language Models are able to change their produced language both structurally and emotionally according to specific prompts eliciting key psychological states (e.g. depression). This means that LLMs are as capable as humans to display alterations of their language in presence of alterations of their thoughts and psychological well-being, a link that goes under the so-called deep lexical hypothesis (Cutler and Condon, 2022): The presence of specific personality traits or psychological constructs can alter language usage and perception, as found in humans affected by depression (Al-Mosaiwi and Johnstone, 2018), or victims of sexual assaults recounting their experiences (Abramski et al., 2024). Whether this link is bidirectional remains an intriguingly open research question but, nonetheless, the ability quantified within CounseLLMe that LLMs can also alter their language in ways similar to humans represents an interesting starting point for future research, opening the way to next-generation and more realistic chatbots for mental health (Eltahawy et al., 2024) but also to more venues of research in understanding LLMs' cognitions via machine psychology (Abramski et al., 2023; Stella et al., 2023).

Our study and dataset come with some limitations. Large Language Models might be dependant on their prompt design and slight edits to prompt engineering might result in different LLMs' behaviour (Zamfirescu-Pereira et al., 2023). To reduce this issue and provide results more consistently related to past approaches, we designed CounseLLMe's prompts in view of the pre-established prompting framework designed by Chen and colleagues Chen et al. (2023), who also studied LLMs' conversations in a mental health setting. Future research could use the CounseLLMe data to quantitatively assess how the emotional trends or the syntactic patterns highlighted here could differ in case of slight changes to the prompts, e.g. using anxiety rather than depression. Another limitation is that the framework of textual forma mentis networks used here, as introduced in Stella (2020) and implemented in Python within EmoAtlas (Semeraro et al., 2025), is powerful in linking together emotions and conceptual associations but lacks the ability to discern between subjects and objects of narratives. In this way, we

cannot understand automatically whether the abnormal levels of sadness expressed by Haiku when impersonating an Italian therapists were due to the model referring to the patient or to itself. A careful human coding of the data revealed that Haiku in Italian interacted with LLaMAntino and the two models together hallucinated, sometimes, in terms of swapping roles despite the reminder prompt. This pattern highlights the crucial importance of accounting for and potentially reducing hallucinations in LLMs working within the mental health domain, since hallucinations can greatly degrade the realism of chatbots.

Fundings

This research received no external funding.

CRediT authorship contribution statement

Edoardo Sebastiano De Duro: Data curation, Formal analysis, Methodology, Validation, Writing – original draft. **Riccardo Improta:** Conceptualization, Formal analysis, Methodology, Software, Validation, Writing – original draft. **Massimo Stella:** Conceptualization, Data curation, Formal analysis, Methodology, Supervision, Validation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge Luciana Ciringione for insightful feedback and Aseem Srivastava for the access to the HOPE dataset.

Data availability

CounseLLMe is publicly available and downloadable from the following Open Science Foundation repository: here.

References

Abramski, K., Ciringione, L., Rossetti, G., Stella, M., 2024. Voices of rape: Cognitive networks link passive voice usage to psychological distress in online narratives. Comput. Hum. Behav., 108266

Abramski, K., Citraro, S., Lombardi, L., Rossetti, G., Stella, M., 2023. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. Big. Data. Cogn. Comput. 7 (3), 124.

Al-Mosaiwi, M., Johnstone, T., 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clin. Psychol. Sci. 6 (4), 529–542.

Alanazi, A.A., Nicholson, N., Thomas, S., 2017. The use of simulation training to improve knowledge, skills, and confidence among healthcare students: a systematic review. Internet. J. Allied. Health. Sci. Pract. 15 (3), 2.

Basile, P., Musacchio, E., Polignano, M., Siciliani, L., Fiameni, G., Semeraro, G., 2023. Llamantino: Llama 2 models for effective text generation in italian language. arXiv preprint arXiv:2312.09993.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J., 1961. An inventory for measuring depression. Arch. Gen. Psych. 4 (6), 561–571.

Bertolazzi, L., Mazzaccara, D., Merlo, F., Bernardi, R., 2023. ChatGPT's information seeking strategy: Insights from the 20-questions game. Proceedings of the 16th International Natural Language Generation Conference, pp. 153–162.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. 33, 1877–1901.

Cece, V., Guillet-Descas, E., Nicaise, V., Lienhart, N., Martinent, G., 2019. Longitudinal trajectories of emotions among young athletes involving in intense training centres: Do emotional intelligence and emotional regulation matter? Psychol. Sport Exerc. 43, 128–136.

Chen, S., Wu, M., Zhu, K.Q., Lan, K., Zhang, Z., Cui, L., 2023. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. arXiv preprint arXiv:2305.13614.

Cheng, S.-W., Chang, C.-W., Chang, W.-J., Wang, H.-W., Liang, C.-S., Kishimoto, T., Chang, J.P.-C., Kuo, J.S., Su, K.-P., 2023. The now and future of chatGPT and GPT in psychiatry. Psych. Clin. Neurosci. 77 (11), 592–596.

Chiu, Y.Y., Sharma, A., Lin, I.W., Althoff, T., 2024. A computational framework for behavioral assessment of LLM therapists. arXiv preprint arXiv:2401.00820.

Cho, Y.-M., Rai, S., Ungar, L., Sedoc, J.a., Guntuku, S.C., 2023. An integrative survey on mental health conversational agents to bridge computer science and medical perspectives, 2023, 11346.

Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., D'Alfonso, S., 2023. To chat or bot to chat: Ethical issues with using chatbots in mental health. Digit. Health. 9, 20552076231183542.

Colby, K.M., Weber, S., Hilf, F.D., 1971. Artificial paranoia. Artif. Intell. 2 (1), 1–25.

Cutler, A., Condon, D.M., 2022. Deep lexical hypothesis: Identifying personality structure in natural language. J. Pers. Soc. Psychol.

Demasi, O., Li, Y., Yu, Z., 2020. A multi-persona chatbot for hotline counselor training. Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3623–3636.

D'Alfonso, S., 2020. Ai in mental health. Curr. Opin. Psychol. 36, 112–117.

Eltahawy, L., Essig, T., Myszkowski, N., Trub, L., 2024. Can robots do therapy?: Examining the efficacy of a CBT bot in comparison with other behavioral intervention technologies in alleviating mental health symptoms. Comput. Hum. Behav.: Artificial Humans 2 (1), 100035.

Fatima, A., Li, Y., Hills, T.T., Stella, M., 2021. Dasentimental: Detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning. Big. Data. Cogn. Comput. 5 (4), 77.

Fiske, A., Henningsen, P., Buyx, A., 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J. Med. Internet Res. 21 (5), e13216.

Fitzpatrick, K.K., Darcy, A., Vierhile, M., 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. JMIR Ment. Health. 4 (2), e7785.

Grodniewicz, J.P., Hohol, M., 2023. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. Front. Psychiatry. 14, 1190084.

Hagendorff, T., 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv preprint arXiv:2303.13988.

Haim, E., Stella, M., 2023. Cognitive networks for knowledge modelling: A gentle tutorial for data-and cognitive scientists dsmuk.

Hinkle, D.N., 1965. The change of personal constructs from the viewpoint of a theory of construct implications. The Ohio State University.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Holtzman, N.S., et al., 2017. A meta-analysis of correlations between depression and first person singular pronoun use. J. Res. Pers. 68, 63–68.

Inkster, B., Sarda, S., Subramanian, V., et al., 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR mHealth and uHealth 6 (11), e12106.

Kian, M.J., Zong, M., Fischer, K., Singh, A., Velentza, A.-M., Sang, P., Upadhyay, S., Gupta, A., Faruki, M.A., Browning, W., et al., 2024. Can an LLM-powered socially assistive robot effectively and safely deliver cognitive behavioral therapy? a study with university students. arXiv preprint arXiv:2402.17937.

Lambert, M.J., Gregersen, A.T., Burlingame, G.M., 2004. The outcome questionnaire-45.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G., 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. 55 (9), 1–35.

Lovibond, P.F., Lovibond, S.H., 1995. The structure of negative emotional states: Comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. Behav. Res. Ther. 33 (3), 335–343.

Malhotra, G., Waheed, A., Srivastava, A., Akhtar, M.S., Chakraborty, T., 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. Proceedings of the fifteenth ACM international conference on web search and data mining, pp. 735–745.

Melo, A., Silva, I., Lopes, J., 2024. ChatGPT: A pilot study on a promising tool for mental health support in psychiatric inpatient care. Int. J. Psychiatric Trainees.

Meskó, B., 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. J. Med. Internet Res. 25, e50638.

Mohammad, S.M., Turney, P.D., 2013. Crowdsourcing a word–emotion association lexicon. Comput. intell. 29 (3), 436–465.

Moriya, R., 2019. Longitudinal trajectories of emotions in four dimensions through language advisory sessions. Studies in Self-Access Learning Journal 10 (1).

Pham, K.T., Nabizadeh, A., Selek, S., 2022. Artificial intelligence and chatbots in psychiatry. Psychiatr. Q. 93 (1), 249–253.

Plutchik, R., 1991. The emotions. University Press of America.

Pyszczynski, T., Holt, K., Greenberg, J., 1987. Depression, self-focused attention, and expectancies for positive and negative future life events for self and others. J. Pers. Soc. Psychol. 52 (5), 994.

Rai, S., Stade, E.C., Giorgi, S., Francisco, A., Ungar, L.H., Curtis, B., Guntuku, S.C., 2024. Key language markers of depression on social media depend on race. Proc. Natl. Acad. Sci. 121 (14), e2319837121.

Raz, T., Luchini, S., Beaty, R., Kenett, Y., 2024. Automated scoring of open-ended question complexity: A large language model approach.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E.M., et al., 2020. Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637.

Sedlakova, J., Trachsel, M., 2023. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am. J. Bioeth. 23 (5), 4–13.

Semeraro, A., Vilella, S., Improta, R., et al., EmoAtlas, 2025. An emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science. Behav. Res. 57, 77. https://doi.org/10.3758/s13428-024-02553-7.

Song, I., Pendse, S.R., Kumar, N., De Choudhury, M., 2024. The typing cure: Experiences with large language model chatbots for mental health support. arXiv preprint arXiv:2401.14362.

Spring, T., Casas, J., Daher, K., Mugellini, E., Abou Khaled, O., 2019. Empathic response generation in chatbots. Proceedings of 4th Swiss Text Analytics Conference (SwissText 2019), 18-19 June 2019, Wintherthur, Switzerland. 18-19 June 2019.

Stapleton, L., Taylor, J., Fox, S., Wu, T., Zhu, H., 2023. Seeing seeds beyond weeds: green teaming generative AI for beneficial uses. arXiv preprint arXiv:2306.03097.

Stella, M., 2020. Text-mining forma mentis networks reconstruct public perception of the STEM gender gap in social media. PeerJ Comput. Sci. 6, e295.

Stella, M., 2022. Cognitive network science for understanding online social cognitions: A brief review. Top. Cogn. Sci. 14 (1), 143–162.

Stella, M., Hills, T.T., Kenett, Y.N., 2023. Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. Proc. Natl. Acad. Sci. 120 (43), e2312911120.

Tanana, M.J., Soma, C.S., Srikumar, V., Atkins, D.C., Imel, Z.E., 2019. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. J. Med. Internet Res. 21 (7), e12529.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Adv. Neural. Inf. Process. Syst. 30.

Wang, L., Wang, D., Tian, F., Peng, Z., Fan, X., Zhang, Z., Yu, M., Ma, X., Wang, H., 2021. Cass: Towards building a social-support chatbot for online health community. Proc. ACM Hum. Comput. Interact. 5 (CSCW1), 1–31.

Weizenbaum, J., 1966. Eliza—a computer program for the study of natural language communication between man and machine. Commun. ACM 9 (1), 36–45.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C., 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382.

Wolf, T., Sanh, V., Chaumond, J., Delangue, C., 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:1901.08149.

Zamfirescu-Pereira, J.D., Wong, R.Y., Hartmann, B., Yang, Q., 2023. Why johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–21.